Automatic Initialization for Facial Analysis in Interactive Robotics

Ahmad Rabie[1], Christian Lang[1], Marc Hanheide[1], Modesto Castrillón-Santana[2] and Gerhard Sagerer[1]

[1] Applied Computer Science Group, Fac. of Techn., Bielefeld University, Germany
[2] SIANI, University of Las Palmas de Gran Canaria, Spain
(arabie,clang,mhanheid,mcastril,sagerer)@techfak.uni-bielefeld.de

Abstract. The human face plays an important role in communication as it allows to discern different interaction partners and provides non-verbal feedback. In this paper, we present a soft real-time vision system that enables an interactive robot to analyze faces of interaction partners not only to identify them, but also to recognize their respective facial expressions as a dialog-controlling non-verbal cue. In order to assure applicability in real world environments, a robust detection scheme is presented which detects faces and basic facial features such as the position of the mouth, nose, and eyes. Based on these detected features, facial parameters are extracted using active appearance models (AAMs) and conveyed to support vector machine (SVM) classifiers to identify both persons and facial expressions. This paper focuses on four different initialization methods for determining the initial shape for the AAM algorithm and their particular performance in two different classification tasks with respect to either the facial expression DaFEx database and to the real world data obtained from a robot's point of view.

Keywords: facial analysis, initialization, aam, face detection

1 Introduction

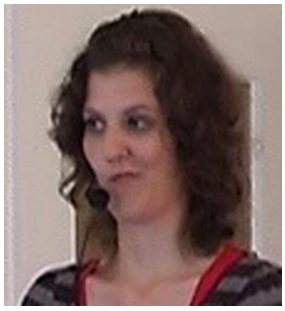
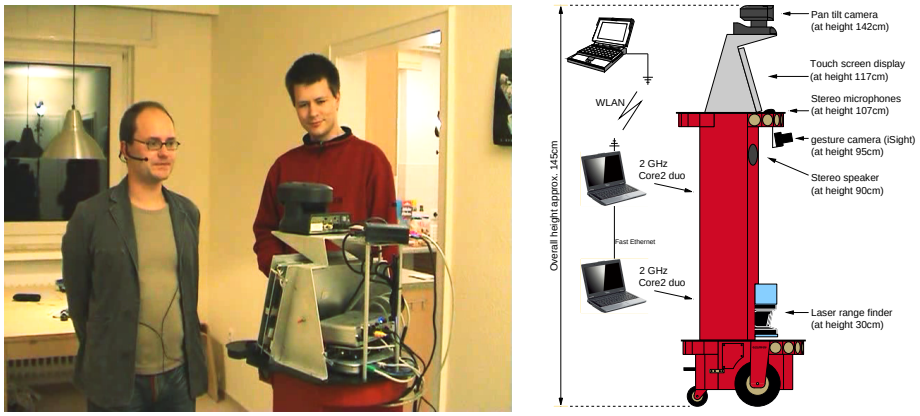


Fig. 1. Facial expression in interaction.

As the face of an interaction partner is one of the most important cues for any interaction, humans evidence very advanced and specialized capabilities to acquire and apply models of human faces. Additionally, when targeting at social robots communicating with humans, the analysis of facial features is a rich source of information for successful and natural interaction. First, the face is considered to be the most discriminant visual feature to identify and discern different interaction partners. This identification is especially important for a robot to provide personalized services and allows for user adaptation in scenarios where it has to cope with several different users as illustrated



(a) Two persons interacting with the robot. (b) Hardware.

Fig. 2. BIRON — the Bielefeld Robot Companion.

in Fig. 2(a) for a typical home environment. Second, the robot has to communicate with a person in a most intuitive and natural way. Communication involves not only verbal but also non-verbal feedback cues that yield information about the human's conversational and emotional state during the interaction with a robot in terms of facial expressions [1]. Especially communication problems such as confusion as illustrated in Fig. 1 can often be read from facial expressions [2] and trigger appropriate dialog and classification behavior in the robot. Consequently, this paper presents a system for visual facial analysis embedded as part of the interactive robot companion BIRON [3]. The robot is enabled to focus on its current interaction partner and to detect other persons of interest. Its person attention mechanism [4] allows it to align its pan-tilt camera accordingly (see Fig. 2(b) for a sketch of the robot), and zoom and focus the face of the respective interaction partner. Its goal is to learn about the environment of its users and to provide personalized services. Besides other components the robot system comprises a speech understanding and dialog system [5] including user modeling and adaptation that can directly benefit from face recognition and facial expression analysis as presented in this paper. In this paper we focus on appropriate detection and initialization methods to account for the real-world challenges of varying view points for facial analysis targeting at emotions recognition and face identification.

Although facial analysis has been well studied in computer vision literature [6, 7] these approaches have mostly been designed for the still image context and rarely for continuous processing [8]. AAMs [9] are one popular solution of the problem of feature extraction from face images. They have been well studied in this domain and also constitute the basic technology for the system presented here. But in the real world domain of social robotics besides the facial feature extraction and analysis also the face detection in the continuous image stream captured from the robot's camera is important. Furthermore, AAMs are based on an iterative optimization scheme that demands an appropriate initialization.

The alignment and initialization for facial analysis using AAMs is studied in this paper and its relevance for applications in the interactive robot companion is evaluated and discussed.

In a related work that discusses the initialization for AAMs, Sattar et al. [10] focused on the face alignment phase for the AAM, but assumed a fixed pose of the face in contrast to the integrated system presented here. Other researchers in human robot interaction have studied facial features mainly for identification of interaction partners. Wong et al. [11] already presented a robot system that is able to recognize humans on the basis of their faces more than ten years ago. However, their approach is limited to six pre-trained faces and rather sensitive to lighting conditions. The humanoid robot ROBITA has also been equipped with a person attention system comprising face recognition [12] to discern different users. Sakaue et al. [13] presented a face recognition system for a dialog interface robot.

In the following we present the general architecture of the facial analysis sub-system. As this paper is focused on the effect and relevance of appropriate initialization a separate section is dedicated to the detection of faces and basic facial features. Afterwards we present basics of the applied AAMs and discuss different types of initialization. A comprehensive evaluation as well on databases as on real world video material unveils the relevance of the developed initialization schemes. Concluding remarks discuss these results.

2 System

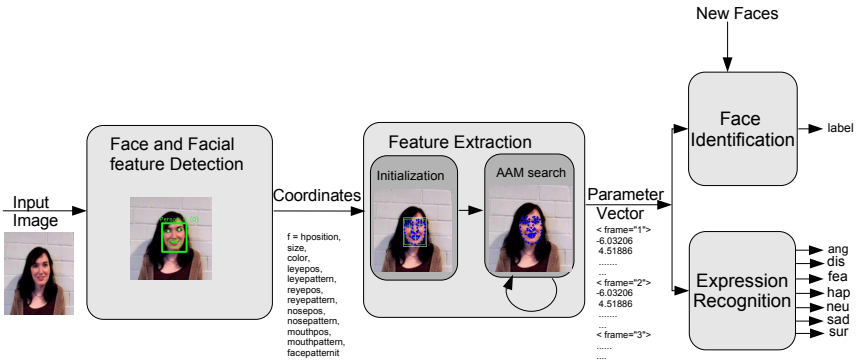


Fig. 3. Architecture of the facial analysis sub-system.

The goal of the facial analysis sub-system as part of the general architecture in BIRON is to (i) discern different interaction partner and (ii) recognize basic emotions that can be used to influence the dialog as one cue of non-verbal feedback [5]. As the basic technique for both goals we apply AAMs. An AAM realizes an iterative optimization scheme, requiring an initialization. Therefore, the AAM fitting algorithm is embedded in a vision system that consists of four basic components as illustrated in Fig. 3. Face pose and basic facial features

(BFFs), such as nose, mouth and eyes, are recognized by the face detection module. The coordinates representing these features are conveyed to the facial feature extraction module. Here, the BFFs are used to initialize the iterative AAM fitting algorithm. A bad alignment of the model drops as well the fitting algorithm as the whole performance of the system [14]. We proposed several methods to initialize the AAM on the basis of detected BFFs and evaluated the results for each of them. After feature extraction the resulting parameter vector for every image frame is either classified to discern different persons or to recognize facial expression related to six basic emotions in addition to the neutral one. Besides the feature vector, AAM fitting also returns a reconstruction error that is applied as a confidence measure to reason about the quality of the fitting and also to reject prior false positives resulting from face detection. For both classifiers, SVMs [15] are applied. However, the requirements derived from the two classification tasks differ for both classifiers. While the identification sub-system is able to learn new faces during the course of interaction when yet unknown persons introduce themselves to the robot, the facial expression classification is trained in advance to work more independently of the person's identity. Both classification schemes can be improved by drawing benefit of the continuous video stream applying a majority voting scheme on the basis of a history of recognition results. The system is applicable in soft real-time, running at a rate of approximate (5) Hz on recent PC hardware.

3 Face Detection

For our facial analysis sub-system, the frontal face and the BFFs have to be detected from the continuous video stream. Face detection has gained the attention of researches in recent years achieving different approaches that solve the problem with great performance [16–18]. However, instead of restricting our approach to a single image based technique as the well known [18], we have preferred an approach that makes use of cue combination to get greater robustness and higher processing speed, particularly for our scenario where live video is processed.

In the face detection approach, a face is initially detected by means of Viola and Jones based detectors: frontal face [19] and upper body [20]. This initial detection allows the system to opportunistically trigger the search of its inner facial details: eyes, nose and mouth. Our assumption is that their detection would improve the precision of the initialization and therefore the AAM search process. Thus once the face has been detected, the facial feature detectors are launched in those areas that are coherent with their expected location for a frontal face. Those located will characterize the face as follows: $f = \langle position, size, color, leye_{pos}, leye_{pattern}, reye_{pos}, reye_{pattern}, nose_{pos}, nose_{pattern}, mouth_{pos}, mouth_{pattern}, face_{pattern} \rangle$

Following this idea, an initial detection provides different features that not only characterize the detected face, but are also used in the next frames to speed up and make more robust the detection process, taking into account the tem-

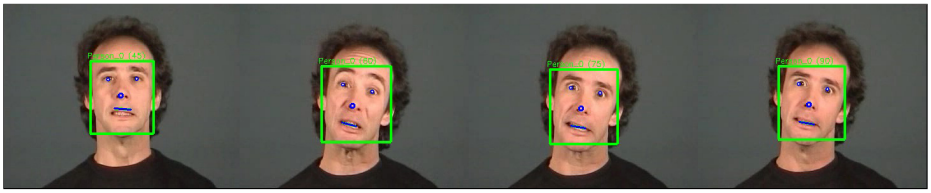


Fig. 4. Face and facial element detection results for some samples of a sequence extracted from DaFEx [21].

poral coherence enclosed in a video stream. Viola and Jones based detection is therefore combined with facial feature tracking and color segmentation, achieving a cue combination approach that in our context provides faster and more robust performance than the single cue based Viola and Jones approach. Some detection results are presented in Fig. 4. Further details can be found in [22].

4 Features Extraction using AAMs

The generative AAM approach uses statistical models of shape and texture to describe and synthesize face images. To build these models, a training set of faces with shape-defining landmarks is required. A linear shape model is obtained by first aligning all landmark sets with respect to scale, translation and rotation and applying a principal component analysis (PCA) afterwards. Thus, a parameter vector can be used to generate a shape. A triangulation algorithm and image warping are applied to transform all images to fit a common reference shape, so that all remaining differences are due to texture variations. A further PCA is applied to get a linear texture model, allowing for texture generation by a parameter vector as in the shape case. Applying a third PCA on the combined parameter vectors yields the appearance model. It can describe and generate both shape and texture using a single appearance parameter vector, which is used as feature vector for the classification. The "active" component of an AAM is a search algorithm that can find the appearance parameter vector representing a new image, given an initial estimation of its shape in terms of the respective parameter vector. This is achieved by evaluation of the gray value differences between the new image and the model-generated texture.

4.1 Initialization

The AAM fitting algorithm requires a suitable initial estimation of the face's shape to find a proper landmark matching (see Fig. 5). In our approach this initialization is based on the detected BFFs (see Sec. 2). Basically we use the mean shape $m = \left(\begin{smallmatrix} m_{x1} \cdots m_{xn} \\ m_{y1} \cdots m_{yn} \end{smallmatrix}\right)^T$ of the AAM as initial shape and place it within the detected face bounding box. The mean shape can be adopted to improve the fitting of the landmarks to the BFFs $f = \left(\begin{smallmatrix} f_{x1} \cdots f_{x4} \\ f_{y1} \cdots f_{y4} \end{smallmatrix}\right)^T$ (centers of right and left eye, nose

and mouth). For each center of such a feature, there is a corresponding landmark in the mean shape. We refer to these special landmarks as *basic landmarks* $p = \begin{pmatrix} p_{x1} \cdots p_{x4} \\ p_{y1} \cdots p_{y4} \end{pmatrix}^T$, whereas all other points of the mean shape are simply called *landmarks*. Fig. 5 depicts the face bounding box as a white rectangle, the BFFs as white crosses, the basic landmarks colored green and all remaining landmarks in blue. Since the detection component will not always robustly find all BFFs, the initialization works flexibly on any partial set given. If, for instance, only the bounding box (no BFFs at all) of the face is detected only a global scaling and positioning is applied. Given detected BFFs, the corresponding basic landmarks are adopted according to one of the following initialization schemes:

- **Linear transformation**: The size and position of the mean shape is linear transformed such that the distance between each BFF and the corresponding basic landmark is minimized: $m' = m \cdot \begin{pmatrix} s_x & 0 \\ 0 & s_y \end{pmatrix} + \begin{pmatrix} d_x \cdots d_x \\ d_y \cdots d_y \end{pmatrix}^T$ where $s_k = \frac{\sum_{i=1}^{4} \sum_{j=i+1}^{4} f_{ki}}{\sum_{i=1}^{4} \sum_{j=i+1}^{4} p_{ki}}$ and $d_k = \frac{1}{4} \sum_{i=1}^{4} f_{ki} - s_k \cdot \frac{1}{4} \sum_{i=1}^{4} p_{ki}$ with $k \in \{x, y\}$
- **Linear warping**: Each basic landmark is moved to fit the corresponding BFF exactly. The surrounding landmarks are also warped, depending on their distance to the BFF and the basic landmark. The displacement decreases linearly to the distance. Formally, for each landmark i, facial feature j and $k \in \{x, y\}$ do: $m'_{ki} = m_{ki} + d_{kij}$ where $r = (m_{xi} - p_{xj})^2 + (m_{yi} - p_{yj})^2$ and $d_{kij} = (f_{kj} - p_{kj}) \cdot (1 - min\{\frac{\sqrt{r}}{w_k}, 1\})$ where w_k is a weight parameter
- **Gaussian warping**: Likewise to the linear warping, but the decrement of the displacement is Gaussian-based: $d_{kij} = (f_{kj} - p_{kj}) \cdot \exp(-\frac{r}{w_k})$

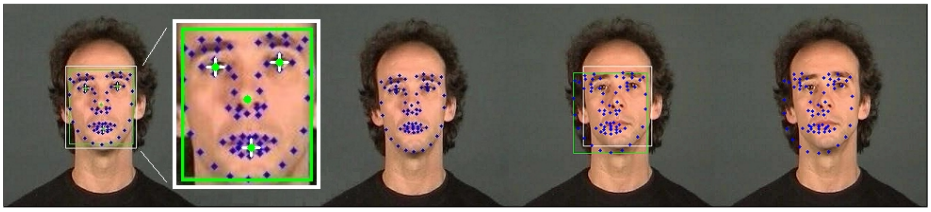


Fig. 5. Initialization based on face bounding box and BFFs (first from left) and landmark matching via AAM search (second) for an image from DaFEx [21]. In cases where the initialization is too poor (third), the AAM search algorithm cannot eventually find a correct matching (fourth).

5 Classification and Evaluation

In the following we present results from evaluation studies carried out with the presented system to assess the appropriateness of the different initialization

techniques. For the evaluation a public available database has been chosen to show the general applicability of the system, furthermore data captured from the view point of our robot is analyzed. The chosen DaFEx database [21] comprises videos of different actors performing different emotions either silently or while talking. The robot corpus has been recorded from the robot's point of view and is applied for the identification sub-task. For all evaluation the AAM has been trained in advance to cover most of the variations occurring between individuals and different facial expressions. Basically, a generic AAM can be trained which is not specific to the dedicated classification tasks presented in the following. All results discussed in the following are obtained on a per frame basis, neglecting the positive effect of possible majority voting on a history of frames.

5.1 Face Identification

An one-vs-all SVM-classifier with linear kernel is used to recognize the identity of known persons. The AAM was trained with 1,056 images (non-talking) from block three of the DaFEx and covers 95% of the training set variance, resulting in a 24-dimensional feature vector. For the classification we randomly selected 100 images per person from block six (non-talking) as training data and 100 images from block one (talking) as test data. Testing with talking faces has been conducted to account for the application domain of our interactive robot. In addition to the tests with the eight persons of the DaFEx we also used videos of twelve persons captured from the robot's perspective. Again, we randomly selected 100 images of each person for training and test (both with talking subjects). The right part of table 1 reports the results in terms of recognition rates (column "Rates").

Considering the BFFs for initialization reduces the reconstruction error[3] and consequently leads to a better representation of the face, compared to the bounding box initialization. Surprisingly, this does not always lead to better classification results. In the DaFEx case it yields slightly higher classification rates for linear and Gaussian warping, whereas linear transformation performs slightly worse despite the good reconstruction error. Even the classification with initialization by placing the AAM mean shape central in the image (without face detection) yields a rate above 95%, though the reconstruction error is very poor. That is due to the unvarying backgrounds and face positions in the training and test data, which causes the AAM search to "fail in a similar way" in both training and test, resulting in poor, but similar feature vectors. Thus the SVM can discriminate the persons nevertheless, although the feature vectors might be inapplicable in terms of the AAM representation.

Unlike the faces in the DaFEx videos, the faces in the videos captured by the robot are not always centered and may also differ in scaling due to different distances. Therefore an initialization by simply centering is not applicable at all. As for the DaFEx videos, Gaussian warping performs best and linear transformation

[3] Sum of squared pixel intensity differences between input image and AAM generated image, column "Rec. Err" in table 1.

	Facial Expression				Face Identification			
	Indiv Model		Gen Model		DaFEx		Robot	
	Rates	Rec Err	Rates	Rec Err	Rates	Rec Err	Rates	Rec Err
Centering	25.70	0.2124	25.31	0.1467	95.38	0.4729	-	-
Bounding Box	71.90	0.0489	67.09	0.0345	99.25	0.0275	95.42	0.1289
Linear Transform	84.30	0.0540	79.04	0.0320	99.00	0.0255	93.25	0.1234
Linear Warping	85.87	0.0485	81.17	0.0186	99.62	0.0236	91.75	0.1180
Gaussian Warping	88.70	0.0472	80.80	0.0180	99.75	0.0224	96.50	0.1104

Table 1. Classification rates and reconstruction errors obtained considering the initialization methods described in section 4.1

yields classification rates poorer than bounding box initialization. Surprisingly, linear warping yields the worst results though not dramatically bad. Also using different kernels (polynomial, RBF) or an one-vs-one SVM did not improve the performance. Tests with another AAM trained with 333 images of 65 subjects from the Spacek database [23] confirmed the poor performance of linear transformation, but yield better results for linear warping. The classification rates ranked usually between those of bounding box initialization and Gaussian warping, which performed the best in this case, too. However, the differences are minor in most cases and not significant, although Gaussian warping tends to perform best.

5.2 Facial Expression Recognition

In order to evaluate the facial expression recognition DaFEx has been used to train and test the system. The robot data set currently does not contain images of different facial expression. Accordingly, no evaluation of expression recognition is carried out on this data. The third block with non talking video data of each actor from DaFEx is selected to train an individual AAM for each actor and also a generic one covering data from all actors. All AAMs are built covering 99% of the training set variance. The parameter vectors of training data of each actor and of all actors are extracted by using the corresponding AAM and are subsequently conveyed to train support vector machine classifier to perform person-dependent and -independent classification into the seven emotion classes. In both person-dependent and -independent cases an one-vs-all SVM-classifier with RBF kernel is used to evaluate the impact of the different initialization methods on the efficiency of the facial expression recognition subsystem.

Table 1 (left part) indicates that the classification rates using the individual models are better than using generic one. That shows the advantage of individual AAMs although their reconstruction errors are larger than those resulting from the generic one. The reason behind is that the variation of the facial features relevant to the expression of one person are smaller than those of multi-person and the classes of individual models are clustered more compact than of the generic one. The smaller reconstruction error of the generic AAM is expected because a larger train data is used in its constructing than in the individual ones. The largest reconstruction errors and the lowest recognition rates occurred by

aligning the model on about the image center (Centering). Coarsely initializing by using the bounding box provided already considerable enhancement of the performance. Minimizing the distance between the facial features and the feature points by using linear transformation initialization offered more adequate AAM fitting and therefore yields better classification results. Moving the basic landmarks and their surrounding to fit the BFFs according to either linear warping or Gaussian warping led to the best performance of the system.

6 Conclusions and Outlook

We presented an integrated vision system for facial analysis and focused on different initialization schemes for AAMs. Our intention was to perform face recognition in the context of human robot interaction to identify the user and also to classify facial expressions occurring in conversation. The presented results show that the information related to the BBFs and especially the way in which they are used improves the AAM fitting process and in consequence the classification performance. The facial expression recognition profits more by the initialization according to the BFFs because the former depends more on the shape recognition, whereas the texture is more important for the latter. The reported results have been achieved by processing single images and should be easily improvable by taking into account the temporal coherence in video. Some authors have used anchoring to combine different modalities in that sense.

In the outlook we expect that sending the information obtained by the AAM fitting process back to the face detector can improve the detection. Furthermore, the classification results have evidenced the benefits of using a person-specific classification model for facial expression recognition. It shall be investigated how an integrated approach directly incorporating the results of identity recognition can improve the recognition of facial expressions. However, this demands online trainable expression models which are therefore in focus of future work similarly as currently already applied for the face identification. Also the surprising result that a good reconstruction error can be paired with low classification rates is worthy to be investigated.

References

1. Darwin, C., Ekman (Editor), P.: The Expression of the Emotions in Man and Animals. 3rd edn. Oxford University Press (1998)
2. Barkhuysen, P., Krahmer, E., Swerts, M.: Problem detection in human-machine interactions based on facial expressions of users. Speech Communication **45**(3) (2005) 343–359
3. Haasch, A., Hohenner, S., Hwel, S., Kleinehagenbrock, M., Lang, S., Toptsis, I., Fink, G.A., Fritsch, J., Wrede, B., Sagerer, G.: Biron – the bielefeld robot companion, Stuttgart, Germany, Fraunhofer IRB Verlag (May 2004) 27–32
4. Fritsch, J., Kleinehagenbrock, M., Lang, S., Plötz, T., Fink, G.A., Sagerer, G.: Multi-modal anchoring for human-robot-interaction. Robotics and Autonomous Systems **43**(2–3) (2003) 133–147

5. Li, S., Wrede, B.: Why and how to model multi-modal interaction for a mobile robot companion. In: AAAI Technical Report SS-07-04: Interaction Challenges for Intelligent Assistants, Stanford, AAAI Press (2007) 71 – 79
6. Chellappa, R., Wilson, C., Sirohey, S.: Human and machine recognition of faces: A survey. Proceedings IEEE **83**(5) (1995) 705–740
7. Zhao, W., Chellappa, R., Phillips, P.J., Rosenfeld, A.: Face recognition: A literature survey. Association for Computing Machinery **35**(4) (2003) 399–458
8. Phillips, P.J., Flynn, P.J., Scruggs, T., Bowyer, K.W., Chang, J., Hoffman, K., Marques, J., Min, J., Worek, W.: Overview of the face recognition grand challenge. In: IEEE Conference on Computer Vision and Pattern Recognition 2005. (2005)
9. Cootes, T.F., Edwards, G.J., Taylor, C.J.: Active appearance models. PAMI **23**(6) (June 2001) 681–685
10. Sattar, A., Aidarous, Y., Gallou, S.L., Seguier, R.: Face alignment by 2.5d active appearance model optimized by simplex. In: ICVS. (2007)
11. Wong, C., Kortenkamp, D., Speich, M.: A mobile robot that recognizes people. In: Proc. Int. Conf. on Tools with Artificial Intelligence, Washington, DC, USA, IEEE Computer Society (1995) 346
12. Matsusaka, Y., Tojo, T., Kubota, S., Furukawa, K., Tamiya, D., Fujie, S., Koabyashi, T.: Multi-person conversation via multi-modal interface: A robot who communicate with multi-user. In: Proc. Eurospeech. (1999) 1723–1726
13. Sakaue, F., Kobayashi, M., Migita, T., Shakunaga, T.: A real-life test of face recognition system for dialogue interface robot in ubiquitous environments. In: ICPR '06: Proceedings of the 18th International Conference on Pattern Recognition, Washington, DC, USA, IEEE Computer Society (2006) 1155–1160
14. Huang, X., Li, S.Z., Wang, Y.: Statistical learning of evaluation function for ASM/AAM image alignment. In: BioAW04. (2004) 45–56
15. Vapnik, V.: The nature of statistical learning theory. Springer, New York (1995)
16. Li, S.Z., Zhu, L., Zhang, Z., Blake, A., Zhang, H., Shum, H.: Statistical learning of multi-view face detection. In: European Conference Computer Vision. (2002) 67–81
17. Schneiderman, H., Kanade, T.: A statistical method for 3d object detection applied to faces and cars. In: IEEE Conference on Computer Vision and Pattern Recognition. (2000) 1746–1759
18. Viola, P., Jones, M.J.: Robust real-time face detection. International Journal of Computer Vision **57**(2) (May 2004) 151–173
19. Lienhart, R., Kuranov, A., Pisarevsky, V.: Empirical analysis of detection cascades of boosted classifiers for rapid object detection. In: DAGM'03, Magdeburg, Germany (September 2003) 297–304
20. Kruppa, H., Castrillón Santana, M., Schiele, B.: Fast and robust face finding via local context. In: Joint IEEE Internacional Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance (VS-PETS). (October 2003) 157–164
21. Battocchi, A., Pianesi, F., Goren-Bar, D.: Dafex, a database of kinetic facial expression. In: ICMI05 Doctoral Spotlight and Demo Proceedings. (2005) 49–51
22. Castrillón Santana, M., Déniz Suárez, O., Hernández Tejera, M., Guerra Artal, C.: ENCARA2: Real-time detection of multiple faces at different resolutions in video streams. Journal of Visual Communication and Image Representation (April 2007) 130–140
23. Spacek, L.: "Collection of Facial Images" WWW [Online]. Available: `http://cswww.essex.ac.uk/mv/allfaces/index.html` Sep. 2007.